Discrete Representation Learning for Modeling Imaging-based Spatial Transcriptomics Data

Dig Vijay Kumar Yarlagadda
Memorial Sloan Kettering/ Cornell University
New York City, USA
yarlagad@mskcc.org

Joan Massagué
Memorial Sloan Kettering
New York City, USA
massaguj@mskcc.org

Christina Leslie
Memorial Sloan Kettering
New York City, USA
lesliec@mskcc.org

Abstract

Imaging-based spatial transcriptomics (ST) provides single-transcript-level spatial resolution for hundreds of genes, unlike sequencing-based ST technologies whose resolution is limited to physical capture regions (spots) on slides. Existing methods to identify patterns of interest in imaging-based ST data are built as extensions of single cell analysis methods, mostly ignoring valuable spatial information encoded in the raw imaging data. Here we present a discrete representation learning approach for modeling spatial gene expression patterns in ST datasets. By employing raw coordinates of detected transcripts and positional encoding of cell centroids as inputs, we learn discrete representations using Vector Quantized-Variational Autoencoder (VQ-VAE) to extract multi-scale structures from fluorescence in situ hybridization (FISH) based ST datasets. We demonstrate the usefulness of discrete representations in terms of the quality of embedding of ST data as well as improved performance on downstream tasks for extracting biologically meaningful cellular neighborhoods and spatially variable genes.

1. Introduction

Spatial transcriptomics (ST) technologies have rapidly progressed in recent years, emerging as powerful next generation tools for biomedical research. ST enables the profiling of gene expression patterns within complex tissues, allowing the identification of cell types and expression states within a spatial context. As cells encounter both direct signals from neighboring cells and soluble signals within their local microenvironment, this spatial context is critical for deeper insights into cell identity and function and for unraveling intricate cell-cell interactions in the native tissue context. In particular, FISH-based ST methods such as Multiplexed Error-robust Fluorescence in situ Hybridization (MERFISH) [1], sequential Fluorescence In Situ Hybridization (seqFISH) [2] and 10x Xenium [3] allow simultaneous profiling of several hundred genes with high spatial resolution. By contrast, sequencing-based ST methods provide a transcriptome-wide readout but much more limited spatial resolution, as transcripts are resolved only to the coordinates of fixed capture regions ("spots") [4, 5, 6]. Moreover, compared to single cell RNA sequencing (scRNA-seq), FISH-based ST techniques can provide a more complete representation of cell types that are fragile and rare, which may be lost in conventional tissue dissociation protocols. The rapid increase in the collection of ST data has led to novel computational challenges in exploiting this data for biological discovery. As ST and scRNA-seq technologies yield distinct data distributions and biases, existing methods for analyzing scRNA-seq data are not suitable for processing ST data.

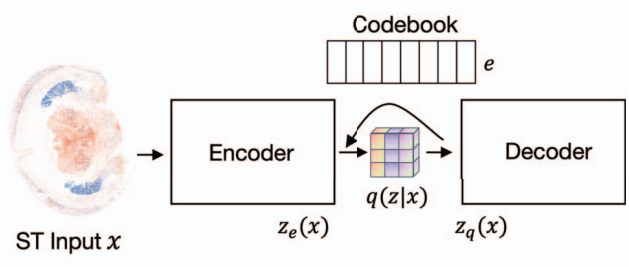

Figure 1. Overview of the proposed discrete representation learning approach for analyzing imaging-based ST data.

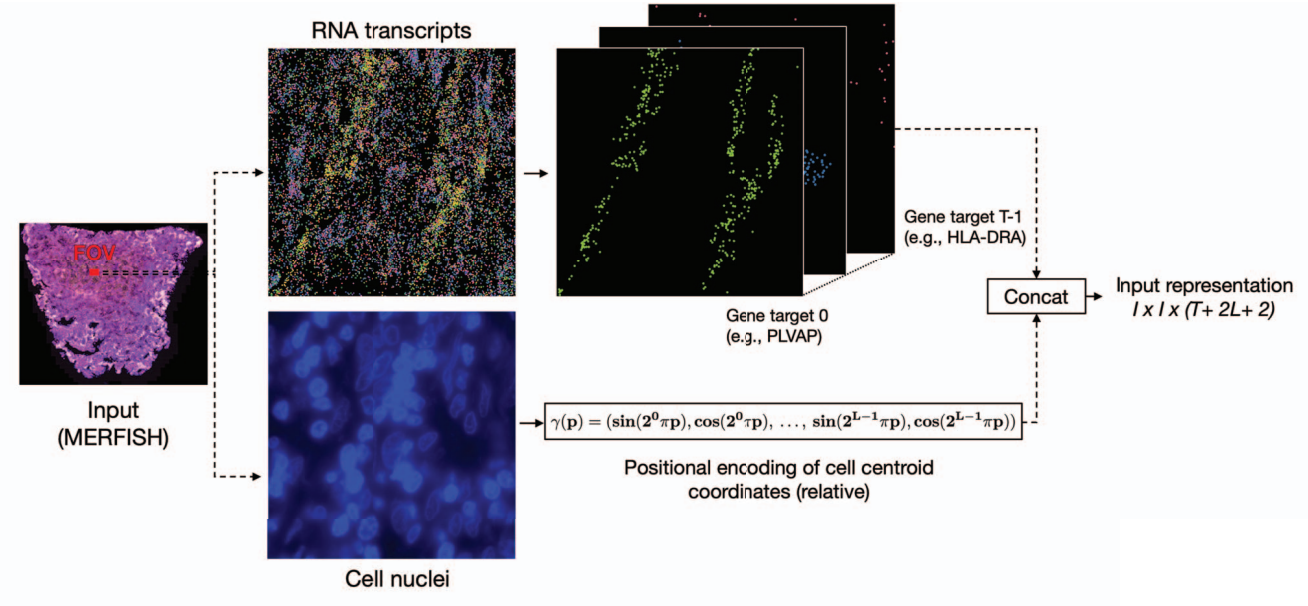

Figure 2. Illustration of input representation for the model. We start with the detected transcripts from imaging-based ST sample across T target channels and concatenate the cell centroid coordinates and positional embeddings of cell centroid coordinates detected from DAPI fluorescence images. Here $\gamma(p)$ represents positional encoding of cell centroid coordinate p, and L represents number of frequencies used for positional encoding.

2. Related work

Several computational frameworks have been developed to preprocess and derive summary statistics from ST samples. Many of these approaches either disregard spatial information altogether or only consider the local spatial context, without the ability to capture broader, long-range spatial information. While dimensionality reduction methods like PCA, t-SNE, and UMAP can be used to visualize and explore high-dimensional spatial gene expression data, there is a need to develop methods that extract information from the spatial distribution of gene expression to uncover tissue- and cell-type-specific gene expression patterns. Several methods have been proposed for this purpose, for tasks such as integration of ST data with scRNA-seq data and identification of spatially variable genes.

Clustering methods are commonly used to identify distinct cellular phenotypes within scRNA-seq and ST data, but they cannot incorporate spatial information captured in ST data. Statistical measures like spatial autocorrelation, Moran's I, and Ripley's statistics have been extensively explored in the literature but found to fall short on ST data, as they require unsuitable assumptions to make comparisons feasible [7, 8, 9]. As the data distribution of scRNA-seq data is overdispersed and contains many zero values due to dropout, it is usually approximated with a zero-inflated negative binomial distribution. Variants on autoencoders [10] such as zero-inflated negative binomial autoencoders [11] have also been introduced to account for the heteroscedastic nature of the transcriptomics data [12, 13]. These methods have been extended to ST data, which is typically modeled with a Poisson distribution [14, 15].

Graph neural networks (GNNs) [16] have also been used to model and analyze the complex interactions and dependencies between spatially proximal cells [17, 18]. First, cell-cell neighbor graphs were built on spatial positions by connecting adjacent cells at a given spatial location when the Euclidean distance of neighboring cells is smaller than 10-30 μm, resulting in each cell connected to 4-8 neighboring cells depending on tissue type, generating the adjacency matrix. GNNs were then applied on this cell-cell graph using a graph autoencoder together with a standard autoencoder to refine the spatial graph structures [18, 17, 19]. Non-negative Matrix Factorization (NMF) is another popular approach to uncover underlying gene modules associated within transcriptomics data and extended to spatial data in non-negative spatial factorization [20] using a Matérn kernel. SpatialDE uses Gaussian processes [21, 22] to identify spatially-variable genes.

In addition, several methods have been proposed to extract neighborhoods of interest in ST data. Similar to cell-cell graphs described above, these methods define neighborhood of a cell as the k nearest neighbors of that cell and generate a gene-gene similarity matrix against mean expression of the reference data. These methods are limited to model a localized spatial context, as they assume that cells are only

Dataset (MERFISH)	Cells	FOVs	Targets
Oncology [23]	8,696,580	207,091	500
Brain [24]	734,705	12,547	483
Hepatocellular [25]	1,671,375	6,700	400

Table 1. Dataset statistics.

affected by a fixed set of k neighbors, and have proven to be ineffective even compared to simpler baselines without these assumptions [18].

The core limitation of these approaches is their heavy reliance on local connections and neighborhood information, potentially missing broader spatial patterns and long-range interactions. Due to the large computational workload required to process imaging-based ST data, often the inputs are heavily downsampled, leading to oversmoothing of gene expression profiles and blurring of finer-scale spatial variations. Moreover, some methods require predefined constraints – such as a fixed neighborhood size, prior knowledge of histological structures, or the need for a reference annotated spatial dataset [17] – which may limit their broader applicability. We focused on overcoming these limitations and developing a robust and scalable approach for capturing both local and global spatial structures while accounting for the unique characteristics of FISH-based ST data.

3. Method

The contributions of our proposed method can be summarized as follows:

- Augmenting the input representation for imaging-based ST data by utilizing the raw spatial coordinates of detected transcripts and positional encoding of cell centroids as model inputs, rather than operating on aggregated cell-by-gene matrices.
- Designing a novel hierarchical encoder tailored for ST data to capture multi-scale spatial structures.
- Leveraging discrete representation learning for extracting biologically meaningful, spatially-informed neighborhoods.

As discrete representation learning has not been previously explored for modeling ST data (to the best of our knowledge), we introduced novel adaptations to VQ-VAE to better capture features of FISH-based ST data. The following sections elaborate these contributions in further detail.

3.1. Input

The primary focus of our proposed method is for analysis of high-resolution, high-throughput cellular-level ST

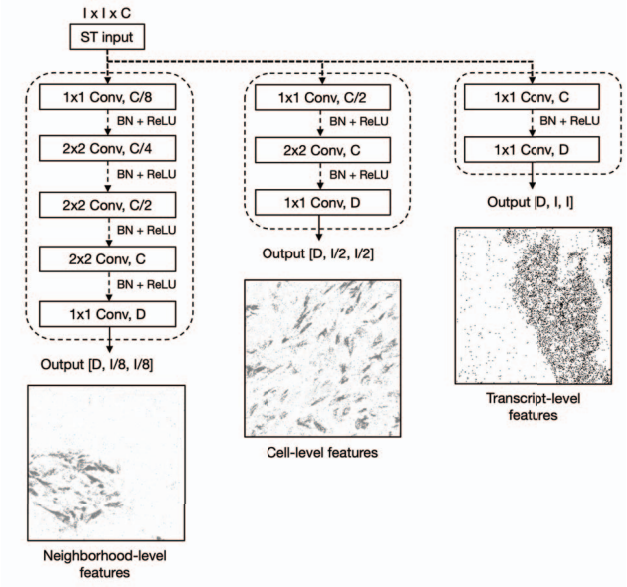

Figure 3. Proposed hierarchical encoder for extracting multi-scale spatial features in ST data. Here, D denotes dimension of codebook embedding, C is number of input channels, I represents height and width of FOV which is set to 256 for MERFISH datasets.

datasets obtained through FISH-based techniques. In FISH-based ST data, pixel values are binary (0 or 1) at the native resolution, denoting the absence or presence of a detected transcript, respectively. Given the high variance nature of ST data, where number of detected gene targets can range from 400 to $10,000$ with sparse single-pixel transcripts for each gene target, many existing methods struggle to effectively model this data. In particular, the level of sparsity is notable compared to natural images: in any channel representing expression for a given target gene, typically less than one percent of pixels contain detected transcripts.

Current methods also struggle to handle the large amount of raw imaging data (approximately 3 TB of gigapixel images per sample) with seven z-planes per field of view, thousands of fields of view per image, and hundreds of thousands of cells. So, they often use summary statistics like cell-by-gene matrices as input to the models. In contrast, we propose to directly use raw coordinates of detected transcripts as input to the model.

We built our model on three diverse MERFISH-based datasets spanning healthy and disease (cancer) tissues: a healthy mouse brain dataset [24], a human oncology dataset containing tissues from 8 distinct cancer types [23] and a hepatocellular carcinoma dataset from patients treated with PD-1 checkpoint blockade [25] (Table 1). These datasets vary in terms of their scales and tissue origins. We first preprocessed the data using the MERLIN image analysis pipeline [29]. MERLIN aligns image stacks obtained from

Method	Params	MSE	SSIM	Scalability
VAE [26]	32M	0.040 ± 0.01	0.36 ± 0.13	1e6
VQ-VAE [27]	3M	0.015 ± 0.003	0.48 ± 0.27	1e7
gimVI [28]	20M	0.160 ± 0.12	0.30 ± 0.02	1e5
NSF [20]	-	0.310 ± 0.02	0.45 ± 0.11	1e4
Ours	19M	$\mathbf{0.003 \pm 0.001}$	$\mathbf{0.824 \pm 0.04}$	**1e9**

Table 2. Test set reconstruction metrics. MSE: lower is better. SSIM: higher is better. Scalability (defined as the maximum number of cells in an input sample that a method can process without exceeding compute requirements of greater than 2 TB of memory or 168 hours of GPU compute per sample): higher is better.

different MERFISH rounds by maximizing their cross-correlation with fiducial bead images and removes background noise with a high-pass filter using blank barcodes as control for non-specific binding. RNA spots are then detected in MERFISH raw data with sub-pixel accuracy using Radial Symmetry-FISH [30] and passed through MERLIN for bit-calling, i.e. decoding a bit as binary 0 or 1 based on fluorescence detection across multiple rounds. The pipeline is run for each imaging field-of-view and then tiled over the entire sample imaging area and z-slices to output the coordinates of detected transcripts.

To approximately assign transcripts to cells, cell segmentation is typically performed using cell segmentation methods such as Cellpose or Baysor [31, 32]. However, certain irregularly shaped cells, for example neurons and macrophages in the brain, will be particularly challenging to segment by using DAPI nuclear fluorescence alone due to their varying soma size and branch architecture. While some methods like MERFISH offer the possibility of protein costaining with cell boundary markers for improved cell segmentation, others like 10x Xenium [3] do not support this feature. So, it remains a challenging problem to accurately assign transcripts to cells.

Inspired by previous work [33] that has reported that information from RNA transcript positions can be leveraged to improve transcript assignment to cells, we chose to use raw detected transcript coordinates as input for our model. Each field of view in the ST data has a native pixel resolution of 100 nm, height and width of I pixels (256 pixels in our dataset), with T channels, each channel representing gene expression for a target gene. We denote the input ST sample as $X \in \mathbb{R}^{W \times H \times T}$, where W and H represent the width and height of tissue capture region respectively (W and H can range from in order of $100,000 - 150,000$ pixels in our dataset). We report the training and test performance (in the Results section) on the MERFISH human oncology dataset [23], which has 16 samples and 207,091 field of views, containing a total of 8,696,580 cells and 4,129,432,299 detected transcripts. We use $S \in \mathbb{R}^{N \times 2}$ to denote the two-dimensional centroid coordinates of each detected cell nuclei in each sample.

To encode spatial relationships between cells, two properties are desirable: (1) distance awareness, to encode pairwise distance between two neighboring cells; and (2) global effectiveness, to encode the relationships between distant spatial structures [34]. We leverage positional encoding (PE) of cell centroid coordinates to encode these properties and enable the model to effectively learn high frequency variation in the ST data. As transformers do not implicity model spatial relationships, positional encoding (PE) was originally proposed [35] for language modeling and later extended for computer vision tasks to embed coordinates of absolute pixel or patch locations in various vision transformer models [36, 37]. Even though convolutional neural networks (CNNs) can efficiently capture spatial information, each field of view in ST data only contains a few hundred cell centroid coordinates corresponding to detected cells. Inspired by NeRF [38], encoding these cell centroid coordinates using positional encoding proved to be critical in embedding the cell-level features into the model.

When using the absolute location coordinates, PE can enable the model to learn the long-range relevance of global relationships between cells, even if they are very far apart. However, this can be a double-edged property as it can weaken the inductive bias of locality, which is often used as a prior to model local cell-cell interactions. Therefore, we use relative positional encodings, reflecting the fact that embedding pair-wise relationships are better suited for capturing cell-cell relationships. Due to the absence of cell boundary staining in modalities like 10x Xenium and the suboptimal cell segmentation performance of current methods, we directly use centroid coordinates of detected cell nuclei rather than the coordinates of cell segmentation boundaries. Our input representation is presented in Fig. 2.

For every cell S_i, we positionally encode the scalar centroid coordinates (x, y) with a sequence of sinusoids with exponentially increasing frequencies [38] as follows:

$$\gamma(p) = (sin(2^0\pi p), cos(2^0\pi p), \ \ldots, \\ sin(2^{L-1}\pi p), cos(2^{L-1}\pi p)) \quad (1)$$

$$\gamma(S_i(x,y)) = (sin(2^0\pi x), sin(2^0\pi y), cos(2^0\pi x), \\ cos(2^0\pi y), \ \ldots, sin(2^{L-1}\pi x), \\ sin(2^{L-1}\pi y), cos(2^{L-1}\pi x), cos(2^{L-1}\pi y)) \quad (2)$$

Here L denotes number of frequencies, which we set to 6 for our setting. We then concatenate the relative position coordinates to this embedding resulting in an input representation of shape $I \times I \times C$, where I represents height and width of FOV which is set to 256 for our data and $C = (T + 2L + 2)$ represents the total number of input channels.

Given this input set-up, our objective is to identify neighborhoods of interest and genes varying within these neighborhoods.

3.2. Model architecture

The overview of our model is presented in Fig. 1. Our model is based on the Vector Quantized Variational Autoencoder (VQ-VAE) architecture [27] and its variants [39, 40, 34], which have proven to be successful in capturing complex spatial patterns in imaging data and compressing this information effectively [41]. Like a standard VAE [26], VQ-VAE is an encoder-decoder architecture. It consists of an encoder that maps an input sample x to a continuous latent space $z_e(x)$, producing continuous latent representations. Unlike the VAE, these latent representations are subsequently fed into a quantizer $q(x)$ to produce a grid of high-dimensional vectors. The quantizer discretizes these continuous representations with a nearest neighbor search by finding the closest embedding vector from a codebook $e \in \mathbb{R}^{K \times D}$, where K is the size of codebook and D is the dimensionality of each discrete latent embedding vector. The output of encoder $z_e(x)$ is mapped to a discrete latent embedding in the codebook as follows:

$$z_q(x) = e_k, \qquad \text{where } k = argmin_j ||z_e(x) - e_j||_2 \quad (3)$$

The size of codebook, i.e., the number of codebook encodings is a hyperparameter and codebook embeddings are also updated during training.

This quantization process allows for the efficient encoding of data as discrete codes and reduces the model's capacity to memorize specific data points. As the gradients used during the forward pass are continuous and the quantization operation is non-differentiable during the backward pass, a straight-through estimator is used to flow through the quantization step during backpropagation. The discrete codes produced by the quantizer are then passed to the decoder, which converts the resultant grid of encodings back into an image. The learned embeddings can be further used with a subsequent model for downstream tasks.

To adapt this model to ST data, we made the following architectural adaptations to the VQ-VAE framework. As

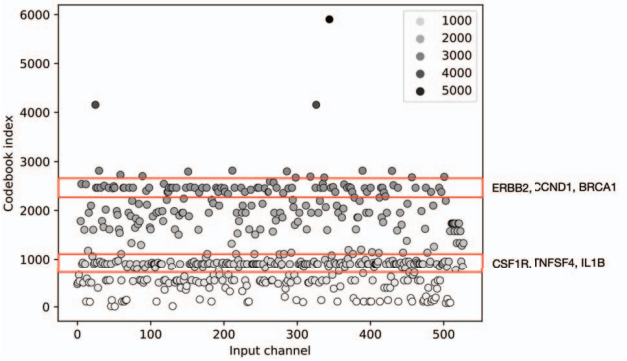

Figure 4. Attribution results. Highlighted genes indicate codebook mappings of cancer-associated genes and macrophage-associated genes respectively, in breast cancer sample from the MERFISH oncology dataset.

individual pixels correspond to detected transcripts in ST data, we observed that often discrete pixels in the input have meaningful biological signal. However, the majority of embedding methods that are currently used for processing ST data are originally designed for photorealism and optimized for reconstruction, which in itself is not the end goal in biology. This results in removal of many transcripts as noise during the reconstruction process, leading to oversmoothing and incorrectly blurred outputs.

We propose to use VQ-VAE, as it can provide a direct correspondence between individual encodings and discrete blocks of pixels in the input image. To address the oversmoothing issue, we first modify the model by utilizing 1×1 convolutional layers. In conventional use of convolutional layers, a common practice is to halve the input size and double the number of filters at each convolutional step. In contrast, our method employs 1×1 convolutions with a stride of 1, which are traditionally used to decrease model complexity. This choice enables the model to initially process the input with fine granularity, considering each detected transcript individually.

Given that the quantization procedure for embedding vectors necessitates a nearest neighbor lookup, an architectural constraint exists such that the dimensionality of the encoder output channels must equal the embedding size D of the learned representations. Since the encoder is constrained to produce an output dimensionality of D, the final convolutional layer must also contain D filters. Therefore, we utilize a 1×1 convolutional layer with a stride of 1 to reduce the number of filters to D while preserving the vector's remaining shape. This also helps to stabilize the training, as it has been also previously been reported that reducing the receptive field size for the convolutions around the relaxation in the VQ-VAE led to it generalizing better to the true evidence lower bound [41]. Thus, a single encoding block consists of 1×1 convolutions as first and last layers and

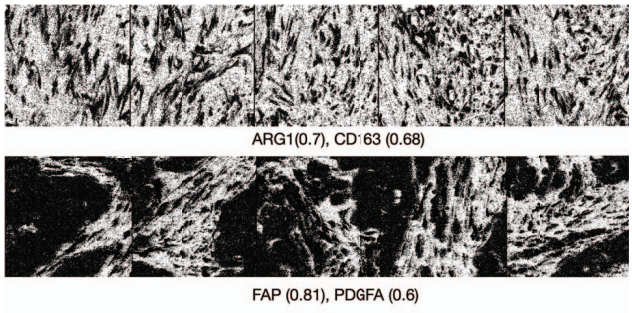

Figure 5. Spatial neighborhoods in the breast cancer sample from MERFISH oncology dataset mapped to distinct codebook vectors in the model (scale, 600 μm). By comparing frequency statistics of mapped encodings, we obtain the ranking of highly expressed genes within the neighborhoods. The neighborhoods in the top panel are rich in genes related to macrophage infiltration, while the neighborhoods in the bottom panel are rich in fibrogenic genes. The morphological similarity of the neighborhoods is notable here, highlighting that the model is able to learn spatial structures corresponding to these different neighborhoods in the tissue and that these can be extracted by inspecting the codebook vectors.

M scaling convolutional blocks in between. Here, each of M convolutional blocks consists of 2×2 convolutional layers with stride of 2, a batch normalization layer, and ReLU activation. Here, hyperparameter M defines the correspondence between pixels and codebook encodings as $I/2^M$.

In ST data, biologically meaningful spatial patterns are observed at multiple scales: at the lowest scale, we can observe the intracellular distribution of gene expression, while cellular level expression can be seen at a progressively higher scale and neighborhood-level information is captured at the highest scale [42, 43, 44]. To encode this information into the model, we designed a hierarchical encoder with three different blocks with varying number of scaling convolutional blocks $M = 0, 1, 3$ to model spatial patterns at progressively higher resolutions as shown in Fig. 3. Using only 1×1 convolutions results in more detail, which we observed is ideal for mapping spatially-variable genes and subcellular structures.

Through empirical evaluation, we determined that the model benefits from additional spatial information capacity during encoding. The decoding process is relatively straightforward, as the contextual information for each cell is encoded within the latent space during the encoding phase. This motivated an asymmetric encoder-decoder design for our model. The decoder starts with a 1×1 convolutional layer to increase the number of filters from discretized encoder output, followed by M transposed convolutional blocks, 1×1 convolution with transposed convolutional layer, and a sigmoid activation to reconstruct the image back to $I \times I \times C$.

3.3. Training

To learn structure awareness, i.e. to encode the structure of the input sample such as density and homogeneity within microenvironment, we additionally use multi-scale spectral similarity index (MSSIM) [45, 46] loss in our model, which is defined as follows:

$$SSIM(x, \hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + c_1)(2\sigma_{x\hat{x}} + c_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2)} \tag{4}$$

where μ_x, $\mu_{\hat{x}}$, σ_x, $\sigma_{\hat{x}}$ and $\sigma_{x\hat{x}}$ are the local means, standard deviations, and cross covariance for input image x and reconstruction $\hat{x}$. Here c_1, c_2 are regularization constants used to avoid instability for image regions where the local mean or standard deviation is close to zero, set to 0.01 and 0.03 respectively.

The multi-scale SSIM is then defined as:

$$MSSIM(x, \hat{x}) = l_R(x, \hat{x})^{\alpha_R} \prod_{j=1}^{R} cs_j(x, \hat{x})^{\beta_j} ss_j(x, \hat{x})^{\gamma_j} \tag{5}$$

where $l_R(x, \hat{x})$ is the luminance comparison; $cs_j(x, \hat{x})$ and $ss_j(x, \hat{x})$ are the contrast and structure comparison at scale j. Here, the exponents α_R, β_j and γ_j adjust the relative importance of luminance, contrast and structural components respectively in a scale-specific manner.

The luminance comparison is given by:

$$l(x, \hat{x}) = \frac{2\mu_{x\hat{x}} + c_1}{\mu_x^2 + \mu_{\hat{x}}^2 + c_1} \tag{6}$$

The contrast comparison is given by:

$$cs(x, \hat{x}) = \frac{2\sigma_{x\hat{x}} + c_2}{\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2} \tag{7}$$

The structure comparison is given by:

$$ss(x, \hat{x}) = \frac{\sigma_{x\hat{x}} + c_3}{\sigma_x\sigma_{\hat{x}} + c_3} \tag{8}$$

where c_1, c_2 and c_3 are regularization constants for the luminance, contrast, and structural terms, which are set to $0.01, 0.03$ and 0.015 respectively.

Additionaly, we use mean squared error (MSE) loss, which is defined as follows for an input image x and reconstruction $\hat{x}$ containing a total of N pixels:

$$MSE(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2 \tag{9}$$

As VQ-VAE defines a uniform prior over z, KL divergence will be constant and can be excluded from the loss.

The total loss for our model can thus be written as:

$$\begin{aligned} Loss = MSSIM(x, \hat{x}) + \|\text{sg}[z_e(x)] - e\|_2^2 + \\ \kappa\|z_e(x) - \text{sg}[e]\|_2^2 \end{aligned} \tag{10}$$

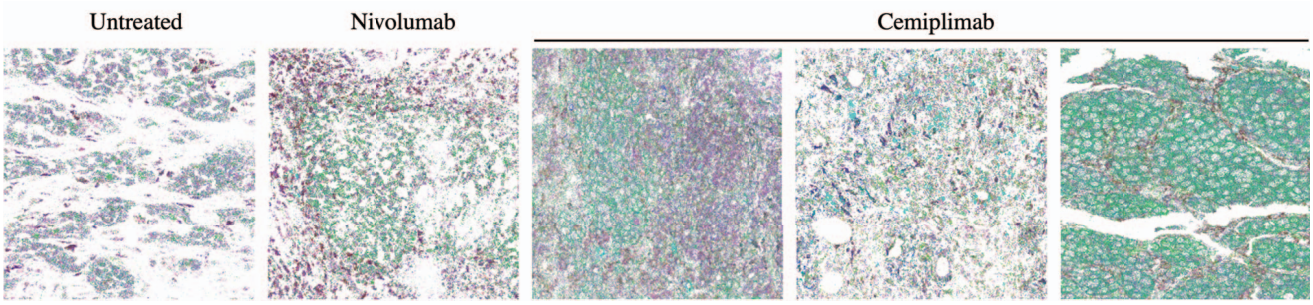

Figure 6. Top spatial neighborhoods (corresponding to most frequently expressed codebook vectors) extracted by the model from patient samples in the hepatocellular carcinoma dataset. Here, all 400 detected transcripts are plotted onto the neighborhoods (colored with histogram equalization for visual clarity, individual transcripts not shown). Differential expression between neighborhoods (with 400 detected targets in the original input sample) revealed enrichment of carcinoma-associated genes (RHOB, CCL21, PROX1) in neighborhood from the untreated sample, immune-related genes (STAT3, FOS, IL6ST) in the nivolumab-treated sample and collagen genes (COL1A1, FN1, COL3A1) in the cemiplimab-treated samples. With the exception of the rightmost sample (where the signal is dominated by high expression of collagen genes), the model is able to identify informative neighborhoods containing a mixture of carcinoma and infiltrating immune cells.

Here sg refers to stop-gradient operator, which prevent gradients from flowing through through encoder (in second term of eq. 10) or quantization module e (in third term of eq. 10) during backpropagation. The third term of eq. 10 corresponds to commitment loss, which acts as a regularizer for the encoder, forcing it to produce representations that are close to the nearest embedding vectors in a codebook of size K (set to 8192) and improve codebook usage; hyperparameter κ is the commitment cost, which is set to 0.25 in our model. We employ an exponential moving average scheme to update codebook embeddings: instead of updating the codebook directly with encoder outputs from each batch, the codebook is updated using a moving average of past embeddings: $e_i = \alpha * e_i + (1-\alpha) * z_i$, where decay parameter α is set to 0.99. This moving average smooths out the fluctuations in the codebook updates and retains the memory of past data when updating the codebook, thereby stabilizing the training.

We use a straight-through estimator for propagating gradients through the quantization operation, which approximates the true gradient of the non-differentiable quantization function $q(x)$ by simply backpropagating the gradient through it. The key idea is to ignore the quantization in the backwards pass, treating $q(x)$ as the identity function during backpropagation.

We performed a 5-fold cross validation on 12 training samples with 4 samples set aside for the test set with the human oncology MERFISH dataset. For augmenting input data during training, we perform random horizontal/vertical flips and random rotations. We update the parameters using AdamW [47] optimizer with $\beta_1 = 0.9, \beta_2 = 0.96, \epsilon = 10^{-8}$, and weight decay multiplier 4.5×10^{-2} and a decay coefficient of 0.99. We use batch size of 16 and and train model to convergence for 30 epochs on four 80 GB NVIDIA A100 GPUs for a total of $675,262$ updates. We performed bayesian hyperparameter optimization to select the hyperparameters described above [48]. Further, we are training the model on 55 manually curated FISH-based ST datasets from the literature, making it the largest reported model to date for ST data.

4. Results

To evaluate the effectiveness of discrete representations, we compare it to state-of-the-art spatial and non-spatial methods. Our first baseline is a standard VAE consisting of a mirrored encoder-decoder architecture, each with 3 convolutional blocks. The VAE is trained for convergence for 30 epochs using the AdamW optimizer with learning rate of 1e-4 and a batch size of 16 on the standard training objective (mean squared error + KL-divergence [49]). Our second baseline is gimVI, a VAE model based on scVI [50] originally designed for scRNA-seq data. It uses alternative conditional distributions to tackle technology-specific covariate shift, to map ST data to reference scRNA-seq datasets. Our third baseline is the standard VQ-VAE without the proposed enhancements. Visually, the reconstruction outputs from the VQ-VAE surpass those of the VAE, which lacks detail and appear blurry. As the number of detected transcripts in ST images is very sparse, standard VAE variants often generate only blank images. We suspect that this is due to the much larger output space: for any given pixel there are 2^C possibilities, but the VQ-VAE model has only $K(8192)$ possibilities for any given pixel. We also compared our model against other models developed for ST, but we notably encountered scalability issues with methods that employ GNNs and Gaussian processes [18, 19, 22] even on the smallest mouse brain dataset, due to their large memory requirements and long running time and in the case of NSF

[20], the requirement to run on CPUs.

All deep learning models are evaluated with 5 random seeds, and the average performance is reported, while statistical models are evaluated only once. Through ablation studies, we verified the effectiveness of the enhancements proposed in our model and found the model to be relatively robust to changes in number of discrete latent embedding vectors and size of the codebook (Fig. 3). We evaluate the models using mean squared error (MSE) and structural similarity index (SSIM). MSE provides a measure of the pixel-wise error, measuring the average squared differences between corresponding pixels in the original and reconstructed images (eq. 9); while SSIM compares the structural similarity between original and reconstructed images based on multi-scale statistical comparisons of luminance, contrast, and structural features (eq. 4). We use these metrics as they are complementary: MSE focuses on individual pixels while ignoring overall image structure while SSIM focuses on the structure of the image, quantifying how well models capture higher-order spatial relationships.

Table 2 compares the models in terms of number of parameters and the quantitative metrics described above, showing that our model outperforms other methods in learning spatial structures in the ST data. Furthermore, our model has an average inference time of 124.2 seconds per MERFISH sample (each sample across the three MERFISH datasets contains an average of 444,106 cells per sample) on a single A100 GPU, which enables it to scale to datasets with billions of cells. Compared to the baselines, our model has significantly lower computational overhead through the use of 1×1 convolutions, binary input pixels and avoiding computing expensive metrics such as covariance.

To further inspect the spatial structures learned by our model, we performed model attribution. First, by feeding one gene expression channel at a time to the model and inspecting the resulting mapping of codebook embeddings, we obtained the mapping between the genes and codebook embeddings. As shown in Fig. 4, cancer-associated genes and macrophage-associated genes are mapped to distinct latent embeddings in the codebook. We then use the encoding-pixel correspondence (defined by hyperparameter M) to map codebook embeddings to distinct spatial neighborhoods (blocks of pixels in the input mapped to the M = 3 encoder network) in the input ST sample. The codebook embeddings learned by our model correspond to different spatially-informed neighborhoods in ST data as shown in Fig. 5.

Furthermore, by inspecting the frequency statistics of mapped codebook embeddings within these neighborhoods, we identified a map of spatially variable genes, which are genes that exhibit distinct expression patterns across different spatial locations, as shown in Fig. 5. Importantly, we demonstrate that the discretization only affects the spatial coordinates and the continuous nature of expression data is preserved and thus can be used with standard approaches such as computing differentially expressed genes across the extracted neighborhoods, as shown in Fig. 6. Similarly, gene expression dynamics can be compared across these neighborhoods with standard trajectory inference approaches [51, 52, 53] .

Codebook size	Emb. dim	Val loss	Perplexity
512	64	1.14	4.226
	256	0.0556	4.077
	1024	0.9389	5.887
1024	64	1.226	7.848
	256	0.1371	6.247
	1024	1.229	8.918
8192	64	0.0271	11.841
	256	1.24	11.330
	1024	0.9389	8.982

Table 3. Ablation results on parameters, size of codebook ($K = 512, 1024, 8192$) and dimension of codebook embeddings ($D = 64, 256, 1024$) for our model trained for 30 epochs on MERFISH oncology dataset. Validation loss (lower is better) and log perplexity, a measure of codebook usage (higher is better).

5. Conclusion

In summary, our proposed method tackles key challenges in analyzing FISH-based ST datasets. We introduce a novel discrete representation model to capture multi-scale spatial structures in imaging-based ST data, enabling biologically relevant interpretations of ST data. Discrete representations also offer notable advantages for large language models, which expect tokenized inputs, and are naturally suitable for reasoning tasks, for example, in predicting how a cellular neighborhood might change in response to interventions [54]. The main limitation of our model is that it is fully unsupervised, limiting its applicability to tasks that use prior biological information, such as cell type annotation. We believe that the discrete representations learned by our model can be effectively used with a downstream semi-supervised model for spatially-informed cell type annotation in the future. All code is provided as open source at `https://github.com/digvijayky/hier_vq_vae_st`. Pretrained checkpoints trained on large ST datasets are made freely available at `https://huggingface.co/digvijayky/hier_vq_vae_st`. We hope the community can utilize these pre-trained checkpoints to fine-tune models on custom datasets for biological discovery.

References

[1] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233), 2015.

[2] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751):235–239, 2019.

[3] 10x Xenium in situ platform. `https://www.10xgenomics.com/platforms/xenium`. (Accessed on 07/20/2023).

[4] 10x Visium spatial gene expression platform. `https://www.10xgenomics.com/products/spatial-gene-expression`. (Accessed on 07/20/2023).

[5] Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

[6] Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature biotechnology*, 39(3):313–319, 2021.

[7] Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L Ibarra, Olle Holmberg, Isaac Virshup, et al. Squidpy: a scalable framework for spatial single cell analysis. *bioRxiv*, 2021.

[8] Zhihua Qiu, Shaojun Li, Ming Luo, Shuanggen Zhu, Zhijian Wang, and Yongjun Jiang. Detection of differentially expressed genes in spatial transcriptomics data by spatial analysis of spatial transcriptomics: A novel method based on spatial statistics. *Frontiers in Neuroscience*, 16:1086168, 2022.

[9] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546, 2022.

[10] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985.

[11] Tian Tian, Martin Renqiang Min, and Zhi Wei. Model-based autoencoders for imputing discrete single-cell RNA-seq data. *Methods*, 192:112–119, 2021.

[12] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome biology*, 20(1):296, 2019.

[13] Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020.

[14] Peiyao Zhao, Jiaqiang Zhu, Ying Ma, and Xiang Zhou. Modeling zero inflation is not necessary for spatial transcriptomics. *Genome Biology*, 23(1):118, 2022.

[15] Florin C Walter, Oliver Stegle, and Britta Velten. FISHFactor: a probabilistic factor model for spatial transcriptomics data with subcellular resolution. *Bioinformatics*, 39(5):btad183, 2023.

[16] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.

[17] Maria Brbić, Kaidi Cao, John W Hickey, Yuqi Tan, Michael P Snyder, Garry P Nolan, and Jure Leskovec. Annotation of spatially resolved single-cell data with STELLAR. *Nature Methods*, 19(11):1411–1418, 2022.

[18] David Sebastian Fischer, Mayar Ali, Sabrina Richter, Ali Ertürk, and Fabian J Theis. Graph neural networks learn emergent tissue properties from spatial molecular profiles. *bioRxiv*, pages 2022–12, 2022.

[19] David S Fischer, Anna C Schaar, and Fabian J Theis. Modeling intercellular communication in tissues using spatial graphs of cells. *Nature Biotechnology*, 41(3):332–336, 2023.

[20] F. William Townes and Barbara E. Engelhardt. Nonnegative spatial factorization, 2021.

[21] Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. SpatialDE: identification of spatially variable genes. *Nature methods*, 15(5):343–346, 2018.

[22] Ilia Kats, Roser Vento-Tormo, and Oliver Stegle. SpatialDE2: Fast and localized variance component analysis of spatial transcriptomics. *bioRxiv*, 2021.

[23] Vizgen MERFISH human immuno-oncology data release. `https://vizgen.com/human-ffpe-immunooncology-release-roadmap/`. (Accessed on 07/20/2023).

[24] Vizgen MERFISH mouse receptor map. `https://vizgen.com/data-release-program/`. (Accessed on 01/04/2022).

[25] Assaf Magen, Pauline Hamon, Nathalie Fiaschi, Brian Y Soong, Matthew D Park, Raphaël Mattiuz, Etienne Humblin, Leanna Troncoso, Darwin D'souza, Travis Dawson, et al. Intratumoral dendritic cell–CD4+ T helper cell niches enable CD8+ T cell differentiation following PD-1 blockade in hepatocellular carcinoma. *Nature Medicine*, pages 1–11, 2023.

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[27] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[28] Romain Lopez, Achille Nazaret, Maxime Langevin, Jules Samaran, Jeffrey Regier, Michael I Jordan, and Nir Yosef. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv preprint arXiv:1905.02269*, 2019.

[29] George Emanuel, seichhorn, Rongxin Fang, Hazen Babcock, leonardosepulveda, and timblosser. r3fang/merlin: archive_20220331_resubmission, Mar. 2022.

[30] Ella Bahry, Laura Breimann, Marwan Zouinkhi, Leo Epstein, Klim Kolyvanov, Nicholas Mamrak, Benjamin King, Xi Long, Kyle IS Harrington, Timothée Lionnet, et al. RS-FISH: precise, interactive, fast, and scalable fish spot detection. *Nature Methods*, 19(12):1563–1567, 2022.

[31] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.

[32] Viktor Petukhov, Rosalind J Xu, Ruslan A Soldatov, Paolo Cadinu, Konstantin Khodosevich, Jeffrey R Moffitt, and Peter V Kharchenko. Cell segmentation in imaging-based spatial transcriptomics. *Nature biotechnology*, 40(3):345–354, 2022.

[33] Gabriele Gut, Markus D Herrmann, and Lucas Pelkmans. Multiplexed protein maps link subcellular organization to cellular states. *Science*, 361(6401):eaar7042, 2018.

[34] Hongzhi Wen, Wenzhuo Tang, Wei Jin, Jiayuan Ding, Renming Liu, Feng Shi, Yuying Xie, and Jiliang Tang. Single cells are spatial tokens: Transformers for spatial transcriptomic data imputation. *arXiv preprint arXiv:2302.03038*, 2023.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[37] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[38] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[39] Akash Saravanan and Matthew Guzdial. Pixel VQ-VAEs for improved pixel art representation. *arXiv preprint arXiv:2203.12130*, 2022.

[40] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. *Advances in neural information processing systems*, 32, 2019.

[41] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[42] Zheng Li and Xiang Zhou. BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome biology*, 23(1):168, 2022.

[43] Yichun He, Xin Tang, Jiahao Huang, Jingyi Ren, Haowen Zhou, Kevin Chen, Albert Liu, Hailing Shi, Zuwan Lin, Qiang Li, et al. ClusterMap for multi-scale clustering analysis of spatial gene expression. *Nature communications*, 12(1):5909, 2021.

[44] David Joon Ho, Dig VK Yarlagadda, Timothy M D'Alfonso, Matthew G Hanna, Anne Grabenstetter, Peter Ntiamoah, Edi Brogi, Lee K Tan, and Thomas J Fuchs. Deep multi-magnification networks for multi-class breast cancer image segmentation. *Computerized Medical Imaging and Graphics*, 88:101866, 2021.

[45] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[48] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

[49] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[50] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[51] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology*, 32(4):381, 2014.

[52] Wen Zhou, Mary A Yui, Brian A Williams, Jina Yun, Barbara J Wold, Long Cai, and Ellen V Rothenberg. Single-cell analysis reveals regulatory gene expression dynamics leading to lineage commitment in early T cell development. *Cell systems*, 9(4):321–337, 2019.

[53] Laleh Haghverdi, Florian Buettner, and Fabian J Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998, 2015.

[54] Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David A Ross, Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, et al. SPAE: Semantic pyramid autoencoder for multimodal generation with frozen LLMs. *arXiv preprint arXiv:2306.17842*, 2023.